This paper was presented as part of the main technical program at IEEE INFOCOM 2011

On the Growth of Internet Application Flows: A Complex Network Perspective

Xiaofei Wu*, KeYu
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Beijing, China, 100876.
*: Currently a visiting professor at Stony Brook University
e-mail: {wuxf, yuke}@bupt.edu.cn.

Xin Wang
Department of Electrical and Computer Engineerin
Stony Brook University, Stony Brook, New York, USA.
e-mail: xwang@ece.sunysb.edu

***Abstract*—Internet structure possesses many properties of complex networks. However, existing studies are often constrained to deriving web connections based on partial data collected, and the actual Internet traffic and user behaviors are far from being understood. With detailed traffic flow records collected through powerful hardware-based monitors, we study from the perspective of complex network the characteristics of four types of traffic: P2Pdownload, HTTP, Instant Messaging and overall traffic. Based on the data analysis and comparison of different applications, we confirm that both the distributions of node degree and strength of nodes/edges follow power law but they have significant different exponents. Specifically, taking advantage of the strict timing of the records, we study the dynamics of flow graphs. The growth of edges upon nodes is nonlinear. Edges formed between existing nodes, instead of the ones arriving with new nodes dominate the growth. We also observe linear preferential attachment behaviors in the flow graphs.**

***Index Terms*—-Internet flow; power law; degree distribution; complex network; growth process; preferential attachment.**

I. Introduction

The concept of complex network has been extensively used to represent an abstract idea and a method of investigating the interactions among people and things of our world. Almost all sorts of relations, from the molecular level protein connections, to food web of animals in nature, and relationship of human being such as email and scientific cooperation [1], can be modeled as complex networks. Studies on networks originated from such a large variety of fields have given out surprisingly consistent results, making complex network an active and fruitful research topic [2]. Among the results, the power law feature [3] and small world [4] phenomena are probably the most important findings. The most successful explanation of the power law feature was given by Barabsi and Albert model (BA model) [3], which models the complex network through the growth of edges with preferential attachment of edges to nodes with higher degrees. There are many studies of modifications of the BA model with respect to different kinds of networks [2]. Although the BA model and its variants are successful in theory, the supporting data and observations of the actual growth of a complex network is very difficult to acquire.

The Internet, which may be the largest human engineered infrastructure, is a perfect example of a complex network and has been studied intensively. The physical connections of routers [5], [6], the routing information of BGP and other routing protocols [7], the web page and the hyperlink structure [8], [9] have all shown distinct properties of complex networks, with their important network parameters such as degree distribution, diameter of the network and clustering coefficient all following the complex network models. It is interesting that not only the physical architecture of Internet itself, but also how people using it – the total Internet traffic and website visits, possesses the similar properties of complex networks, with the power law property as its most significant sign [10].

Although there are tremendous interests in complex networks, the earlier studies are mainly based on the manual collection of data on various relationships, so the data sets are normally small, and the observations may be biased. Many recent studies focus on understanding the features of webs, with the connectivity information collected by various crawlers or through web data mining. The information traced may not be complete, and the resulting models or parameters may not reflect the exact connection characteristics. Due to the lack of more complete flow data, the characteristics of global behaviors of web users as well as the actual relation between traffic and user behaviors are far from being well understood. To our best knowledge, there are very limited studies on the Internet behaviors based on actual flow data. In [10], the investigations were made based on sampled data collected from Abilene network (Internet2), which carries only academic and research traffic, and is never congested. Moreover, only those flows involving TCP connections with an endpoint on port 80 were considered as web traffic and studied.

In this paper, complex networks built from detailed Internet traffic flow records are examined. Our flow records were generated by powerful line-speed monitors each with a capturer and a classifier to track the traffic of a 10Gbps trunk link between an access network and the backbone. The records consist of detailed complete traffic information with accurate application classifications. This allows us to more closely observe the interactions among Internet hosts and flows, and more accurately model the characteristics of complex networks. Instead of constraining our work to web connections, we study the user behaviors and transmission characteristics of three major types of Internet traffic, including P2Pdownload, HTTP,

978-1-4244-9921-2/11/$26.00 ©2011 IEEE

and Instant Messaging (IM), as well as the total Internet traffic.

We try to investigate important characteristics of the complex networks derived from the flow data, such as node degree and strength distributions by comparing results from different types of traffic. Particularly, taking advantage of the strict timing information of the records, we analyze the formation process of various complex networks in order to better understand the growth. To the best of our knowledge, we are the first to study the growth and preferential attachment behavior using all the timed records instead of taking only a subset of the records which are sampled from the transmission data, and generate important and quantitative results of various applications. Instead of mainly analyzing the web traffic as done in the literature work, the comparison of connection and traffic features of different types of applications will also provide a guideline for better provisioning of Internet resources.

The main contributions of this paper are as follows:

1) We construct graphs from complete flow records of three different applications, namely P2Pdownload, HTTP, Instant Messaging, and the overall traffic. We analyze these graphs from the perspective of complex networks and results such as the power law distribution are given. We observe a significant difference in degree distribution of different types of traffic. We provide detailed analyses on the results.
2) We examine the growth of flow graphs in a detailed and timely manner. We confirm the linear growth of newly added nodes and edges. On the other hand, the analysis also reveals that the number of edges formed between existing nodes is large and grows nonlinearly.
3) We investigate the preferential attachment processes with flow records. We confirm that the preferential attachment of all applications is linear.

The rest of the paper is organized as follows. In Section 2, we provide a brief review of related work. We introduce our data sources and the flow graph construction procedures in Section 3. We then present and discuss our analysis results in Section 4. Finally, we conclude the paper in Section 6.

II. Related Work

Internet has been studied from the complex network point of view in many aspects. The most basic and straight forward physical structure of routers and links connecting them is observed to follow the power law based on snapshots of Internet topology, with power law exponents of out-degree between 2.15 and 2.48 [6]. On top of the physical links, autonomous system (AS) level connections provide another view of Internet as a complex network. Studies based on BGP and other routing information show that network among ASs also possesses power law and small world properties, while many other important metrics such as joint degree distributions, clustering coefficient, eigenvalue and spectrum properties reveal more detailed properties of Internet [7]. The structures of routers and ASs are already large-scale examples of complex networks, however, they are still impacted by factors such as human engineering or geographic constraints. The huge network of the World Wide Web, consisting of web pages and hyperlinks, is enormous in size and constructed in a totally uncontrolled and distributed manner. The WWW of billions of pages and links presents perfect properties of scale free and small world [9], indicating that the theory of complex network provides a fundamental way of describing how things in our world interact with each other.

The complex network of Internet goes beyond the static scenarios. For example, dynamic traffic and flows of data are created when people surf the web. The dynamic traffic and flows may be of more importance, because they reflect how the web works – carrying information among people. Traffic flows give us another view of Internet, by forming a dynamic and active complex network where power law and other properties are also observed [10]. In [10], flow records captured in Internet2 were analyzed and the strength of edges, which is the number of bytes of a flow, showed power law behavior over several orders. The power-law exponents of the strength range from 1.7 to 2.4, which are different with respect to the types of nodes (i.e., servers or clients) and directions (in or out). The large variation range implies the traffic distribution is highly skewed and fluctuated, which limits the use of traditional statistics such as mean and deviation. The studies in [10], however, were limited to traffic passing through the port 80, and due to the limitation in the traffic snoopers used, only the sampled traffic data were recorded. Also, Internet2 is used between academic organizations and often under-utilized. In this paper, we make detailed analyses of the characteristics of Internet traffic and user behaviors taking advantage of the complete flow records collected by powerful network traffic monitors. We obtain more interesting results by analyzing traffic from various types of applications.

Besides observations of ubiquitous existence of power law behavior in complex networks, people tried to find out the intrinsic reason of this feature. Power law behavior is explained most successfully by the BA model [3], which describes the network growth with two necessary parts, the incremental growth and the preferential attachment. The growth of scientific citation networks and other types of networks are studied to verify BA model and its variants. The growth of WWW was also studied and different models were presented to explain the observed data [11]. However, due to the large scale of these networks, only a few snapshots can be captured to demonstrate the growth in most cases. Different from the snapshots produced by crawler or other static records, the flow data or call detail records (CDR) contain strict timing information and precise traffic conditions, and are thus capable of revealing more detailed growth processes of the flow graphs. However, although complex networks constructed by these data were also reported [10], [12], the timing information was not fully utilized to capture the growth process. One purpose of this paper is to investigate in detail the growth process, and to compare the characteristics of different applications, based on the strict timing information incorporated with the flow records.

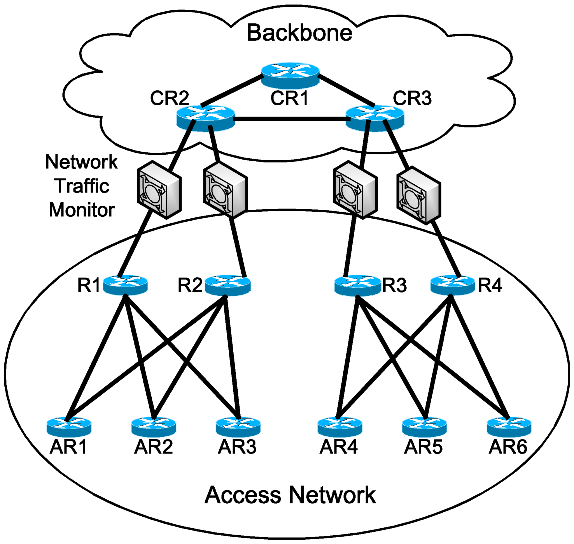


Fig. 1. Network Traffic Monitor and Network

III. Data Sets and Flow Graph

Before presenting our analysis, we introduce the data used in the study of this paper.

Flow data were collected by placing high-performance network traffic monitors on the trunks between an access network (AN) and the backbone. The AN itself is large in scale. It covers a province in southern China, comprises several service systems and Internet Data Centers (IDCs), and serves more than 10 million users. The conceptual diagram of the network architecture is shown in Fig.1.

Each trunk has the capacity of 10Gbps, with its average throughput over 50% and peak traffic over 90% of the total capacity. The network traffic monitor has a high performance hardware probe that captures and classifies every packet passing through the trunk in both directions, which provides continuous flow records. Comprehensive algorithms are used to classify flows into different applications such as Web, FTP, Email, VoIP, Video Stream, P2P, etc. Each flow record contains information about a single network flow, which is defined as one or more packets sent from a source host and port, to a destination host and port, using a particular protocol, and over a certain time interval.

The flow data we use were collected by network traffic monitors in a 24-hour period on March 23, 2010. A total of 1.5 billion flow records were collected with detailed entries including the time stamp, source and destination IP addresses and ports, the total number of packets and bytes in the flow, and the application type in each flow record. Among the more than 15 applications recognized, this paper focuses on P2Pdownload, HTTP, Instant Messaging (IM), which are some of the most typical applications and constitute more than 60% records of the total flow data. We also aggregated all the flow records to construct an overall graph as a reference. The constructed graphs are called flow graphs. Of these traffic types, the P2Pdownload is generally machine-oriented, in the sense that the connections are made by computer programs in a distributed manner, while HTTP and IM traffic are human-oriented which involve human users to decide when and whom to contact and require supports from servers.

TABLE I
BASIC METRICS OF FLOW GRAPHS

Graph	P2Pdownload	HTTP	IM	Overall
N	362057	174449	24858	1209038
E	560861	571237	36210	2482610
$< d >$	3.10	6.56	2.92	4.10
$< d_i >$	1.55	3.28	1.46	2.05
$< d_o >$	1.55	3.28	1.46	2.05
$max(d)$	2065	19614	2304	19820
$max(d_i)$	1328	13868	2192	14017
$max(d_o)$	878	5746	228	5979
σ_d	9050.61	26754.58	3982.61	40827.33
σ_{di}	5414.12	17571.75	3569.27	24841.64
σ_{do}	3946.91	11179.81	994.47	18278.09

Flow records were used to construct flow graphs in the following way: each IP address in the records represents a host or a node; each flow record between two hosts forms an edge; multiple records between the same pair of hosts are considered as the same edge, but the bytes are summed up to give a weight property of the edge, which is called strength of the edge. Since the flows are directional, the flow graphs generated are directed as well.

The flow records we use are far from revealing all the traffic in the network. Actually, only the flows between AN and the backbone were captured, the flows within AN or backbone are not seen by the monitors. So we have no idea how heavy the traffic is within the two parts, and the graph recovered from the data set is far from a complete one. However, our study is still very important. By focusing on the flows crossing the border, we can understand more precisely the interactions between different network segments. In the real network, these are often interactions between autonomous systems, and the interface under monitoring is often the place where billing and traffic control are incurred, and is also the traffic bottleneck in many cases. The analysis in this paper is helpful for better network management and engineering.

IV. Results and Analysis

In this section we present results from the detailed flow records, together with analysis. We begin with degree distribution which illustrates the basic complex network property of the graphs generated from the flow records, and proceed with more results on the properties of complex network growth and connection formation.

A. Degree and basic graph metrics

From the flow records, we construct four flow graphs, namely the P2Pdownload, HTTP, IM and the overall traffic. For HTTP, some properties are further studied for servers and clients separately. Basic metrics are summarized in Table I. In Table

I, N and E are the number of nodes and edges respectively; d, d_i, d_o are degree in total, in-degree and out-degree; $max()$ is the maximum of the corresponding variable, $<>$ stands for the average value and σ is the standard deviation of a variable. For all graphs, $< d >$ is calculated as:

$$< d >= 2E/N \tag{1}$$

$< d_i >$ and $< d_o >$ are statistically counted, i.e. a summation of $< d_i >$ is calculated by adding d_i of every node in a graph and then average over the number of nodes, and $< d_o >$ is obtained similarly. For all graphs, we can see that:

$$< d_i >=< d_o >=< d > /2 \tag{2}$$

which should be exactly like this by definition in a directed graph. This is one evidence that our statistics are correct.

As can be seen in Table I, the average degree values of P2Pdownload and IM are smaller than those of HTTP, and the differences of $max(d)$ between the three applications are about an order, which is significant. This observation may imply that P2P applications are distributed and connected more evenly through the network, and thus probably can use network resources more efficiently. However, this may not be the only reason for their low average degrees, since P2P and IM applications are likely to have more local connections within in AN, which are not captured by our monitor. P2P applications generally favor fast connections which are likely to be local, while in IM applications people that chat might live in the same area therefore being served by the same access network. The HTTP application has the highest $max()$ value, which clearly shows there exist servers acting as hubs in the graph. For all graphs, σ values are at least 3 order larger than those of $<>$, which shows there are really heavy tails in degree distributions. As we know, if the distribution of the degree follows power law with $\gamma > 2$, the second moment does not have a bound. Again the HTTP has larger σ values than those of P2Pdownload. The IM has the lowest average degree values of all graphs, partially because human users are not capable of maintaining as many connections as machines do. We will show more results on the differences between human oriented applications and machine oriented ones in this paper. While these differences are already observed in social and technology networks [17], our data provide more details for this situation. Being different from P2Pdownload where the file can come from any of the file owners, IM traffic often flows between specific users or from the servers. Therefore, although the $< d >$ of IM is roughly the same as P2Pdownload, the $max(d_i)$ of IM is larger than that of P2Pdownload, because central servers exist in IM application. The overall graph contains not only the three applications above, but many more others, so the scale of the overall graph is larger than simply the summation of the three graphs. HTTP has the largest $max()$ values, implying HTTP servers are the most concentrated hosts of Internet.

TABLE II
STATISTICS OF SERVERS AND CLIENTS OF HTTP

	N	$< d >$	$< d_i >$	$< d_o >$	$max(d)$	$max(d_i)$	$max(d_0)$
s	49446	10.85	5.37	5.48	19614	13868	5746
c	125003	4.85	2.45	2.40	1410	666	744

TABLE III
POWER LAW EXPONENTS OF GRAPHS

	P2Pdownload	**HTTP**	**IM**	**overall**	**HTTP_S**	**HTTP_C**
γ_{in}	2.05	1.9	2.15	2.05	1.6	2.14
γ_{out}	2.08	2.05	2.22	2.1	1.75	2.2

For HTTP applications, we further differentiate between server nodes and client nodes, since the roles of clients and servers are totally different, and should show clear differences in statistics. The results were presented in Table II. As expected, the number of servers is much smaller than that of clients, while both the average degree and the maximum degree of servers are significantly larger. The average values of clients are larger than that of P2Pdownload and IM, showing web browsing is the most active application in the Internet. The $max(d_i)$ of servers is much larger than $max(d_o)$, which means that many visits to web sites do not get responses. This observation may imply that in general, servers are overloaded and may have difficulties of providing qualified services. While this is generally acceptable for free web services in most cases, it can be a big problem when considering the quality of service or service level agreement, since when a web site cannot be visited, it is hard for the end users to tell whether it is a server problem which is not the responsibility of an Internet carrier, or a network problem which should not happen since they pay monthly fees.

If the distribution of degree follows the power law, i.e.

$$p(k) \sim k^{-\gamma} \tag{3}$$

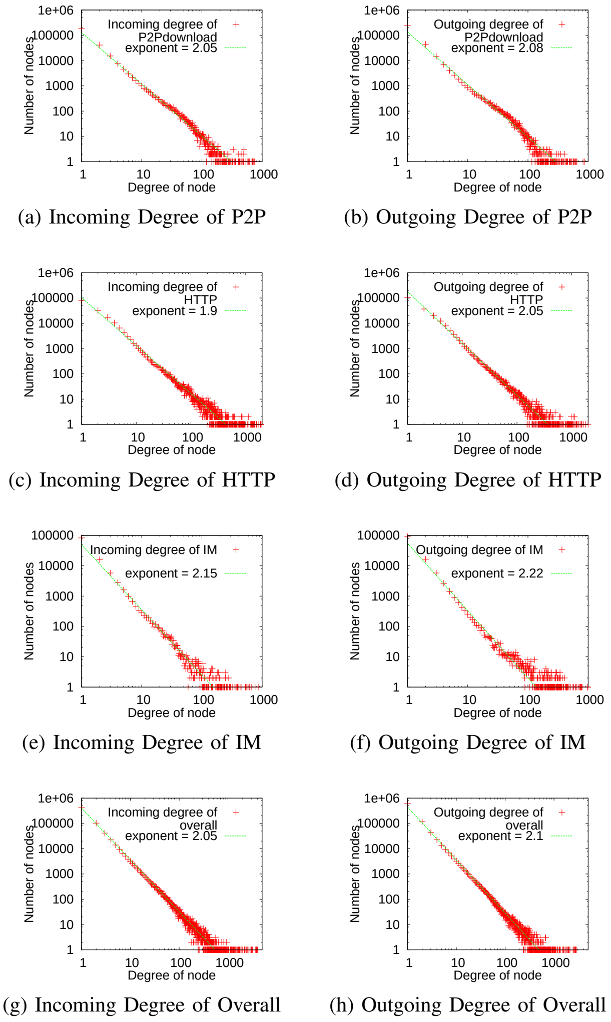
where $p(k)$ is the probability of a node has a degree k, and γ is a constant for a particular network, it can be easily identified on a loglog plot as a straight line. The loglog plots of the degree distributions of the four graphs are shown in Fig.2. The power law is obvious in all graphs, with the γ ranging between 1.6 and 2.25. As expected, all distributions have heavy tails.

For HTTP applications, the degree distributions of servers and clients are also power law. The γ values of all graphs are listed in Table III, where γ_{in} and γ_{out} are power law exponents of incoming and outgoing degree, while HTTP_S and HTTP_C mean HTTP server and HTTP client. We can see that besides the differences in average and max degrees in Table II, the exponent γ of servers and clients are clearly different, with γ of servers smaller than γ of clients. The smaller γ of servers shows that the fraction of high degree nodes of servers is much larger than that of clients.

The graphs of weights are then constructed. The weight of nodes and edges is called *strength*. The strength of an edge is defined as the accumulated flow bytes between two nodes:

$$s_e(i,j) = (bytes\ from\ node\ i\ to\ j) \quad i,j \in [0, n-1] \tag{4}$$

where $s_e(i,j)$ is the strength of edge (i, j). A node has two strengths, an incoming one and an outgoing one, calculated by



(a) Incoming Degree of P2P (b) Outgoing Degree of P2P

(c) Incoming Degree of HTTP (d) Outgoing Degree of HTTP

(e) Incoming Degree of IM (f) Outgoing Degree of IM

(g) Incoming Degree of Overall (h) Outgoing Degree of Overall

Fig. 2. Degree Distributions of flow graphs

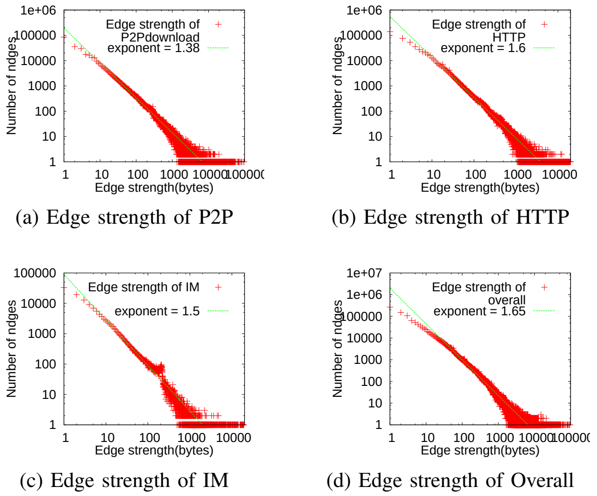
adding all strengths of incoming or outgoing edges of the node, i.e.:

$$si_n(i) = \sum_j s_e(j,i) \qquad (5)$$

$$so_n(i) = \sum_j s_e(i,j) \qquad (6)$$

where $si_n(i)$ is the incoming strength of node i, and $so_n(i)$ is the outgoing strength of node i.

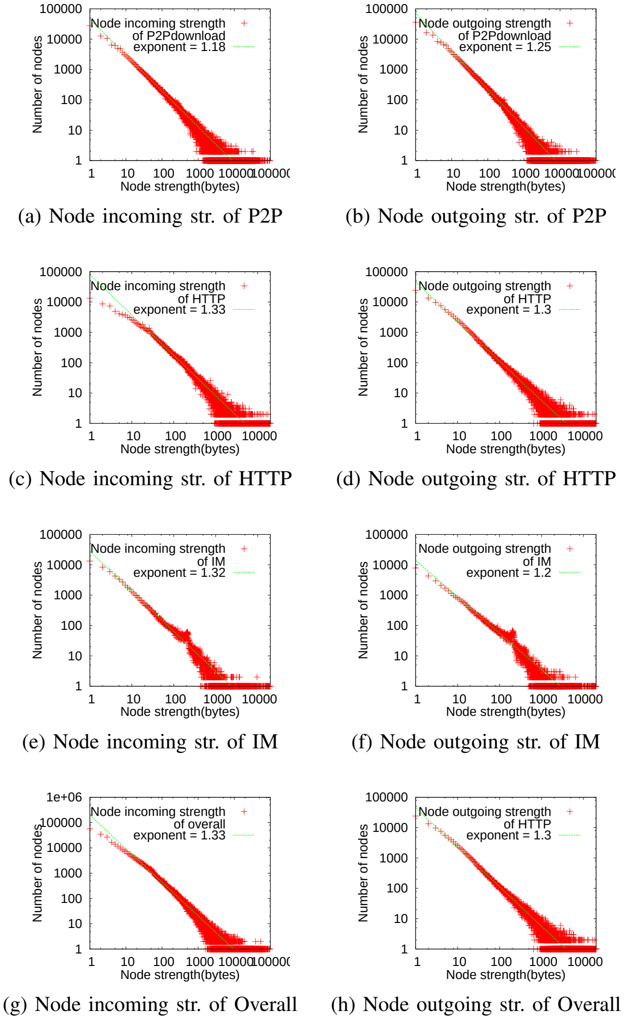
While the degrees of flow graphs show how websites are connected when the Internet is visited, the strengths of nodes and edges reveal the traffic features of the visits, i.e. how many data are transmitted over an edge or how heavily a website is visited. The distributions of strength of aggregated web traffic were studied in [10] and power law feature also holds. We further studied the strength distribution of different applications, and results are presented in Table IV and Fig.3-4. The results in Table IV are interesting. The $< s_e >$ of P2Pdownload is larger than that of HTTP, which is reasonable. The $max(s_e)$ of P2Pdownload is smaller than the $max(s_e)$ of HTTP, and the same situations hold for si_n and so_n. This again gives us the



(a) Edge strength of P2P (b) Edge strength of HTTP

(c) Edge strength of IM (d) Edge strength of Overall

Fig. 3. Edge strength distributions of flow graphs



(a) Node incoming str. of P2P (b) Node outgoing str. of P2P

(c) Node incoming str. of HTTP (d) Node outgoing str. of HTTP

(e) Node incoming str. of IM (f) Node outgoing str. of IM

(g) Node incoming str. of Overall (h) Node outgoing str. of Overall

Fig. 4. Node strength distributions of flow graphs

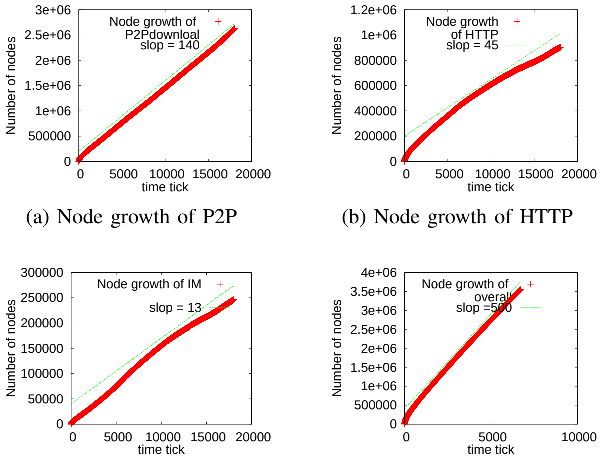
TABLE IV
STATISTICS OF STRENGTHS OF FLOW GRAPH

Graph	P2Pdownload	HTTP	IM	Overall
$< s_e >$	1237865.97	335192.47	69872.83	436320.27
$max(s_e)$	675817801	20135828685	305441256	2013595208
$< si_n >$	1151394.33	1097594.95	101781.93	895928.28
$max(si_n)$	1629060848	3199769428	305537550	3452844589
$< so_n >$	1151394.33	1097594.95	101781.93	895928.28
$max(so_n)$	1663789775	2013582868	305441256	2152802983
γ_{es}	1.38	1.6	1.5	1.65
γ_{nis}	1.18	1.33	1.32	1.33
γ_{nos}	1.25	1.13	1.2	1.45

evidence that the P2Pdownload application is better designed with more balanced traffic transmissions through Internet, thus is more network friendly. The IM application has weights smaller than both P2Pdownload and HTTP. The loglog plots of strength distributions in Fig.3 confirm that the power law exists for all traffic tested. The γ of P2Pdownload (Fig.3a) is smaller than the γ of HTTP (Fig.3b), showing a flatter shape of the power law. This indicates that there is a higher percentage of large weight values for P2Pdownload than that for HTTP, which is the reason for its larger average strength. The strength distribution of the overall graph (Fig.3d) seems to have a consistent bend in the plot, which may imply that the strength of some traffic unknown does not follow the general power law. While the power law behavior is confirmed by our results, the value of γ obtained here is small, ranging from 1.13 to 1.65 compared with 1.7 to 2.4 in [10]. This may be due to the reason that our records are complete, while records in [10] are sampled. Sampled data tend to favor the data transmitted at a higher frequency, thus in our case, favor the strength with smaller values. Due to the extremely skewed power law distribution, favoring samples with smaller strength values could significantly reduce the number of samples with higher strength values, and thus result in a faster decreasing distribution, i.e. a larger γ.

B. Growth of Graph

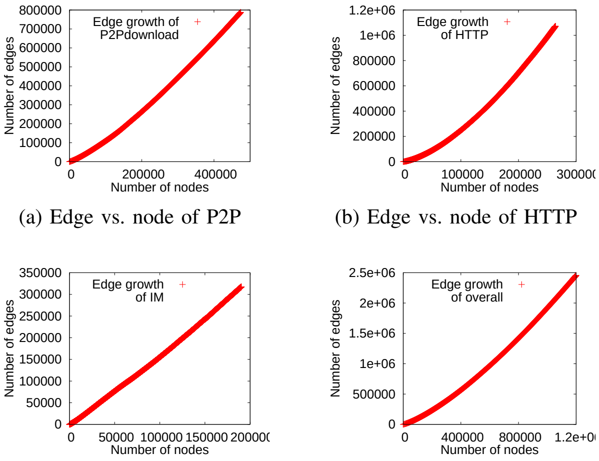
BA model [3] is the most successful one to explain the power law which is commonly used in complex networks. In BA model, new nodes are constantly added to the network, and edges are created between the newly added nodes and existing nodes. The nodes and edges are added linearly, i.e. in each step, a constant number of nodes and edges are added to the graph. The edges are created in a way called preferential attachment [3], which means the probability of a node gets connected by an edge is proportional to the number of edges the node already has. The linear growth of nodes and edges leads to a γ of 3, which is not the case in a practical complex network where most γ vary between 1.2 and 2.5 [2]. In the original BA model, only edges associated with a new node are added, and there are no new edges formed between the nodes already existing, which may not be the case in a practical network. In reality, as in the web case, some web sites are modified constantly, adding and removing links. To compromise these cases, BA model is modified in many ways including considering nonlinear growth [13] and considering birth of new edges between the existing nodes [14]. However, although efforts are made to study the growth models of complex networks, the actual growth is not easy to observe. Our flow graphs, generated based on flow records with strict timing information, give us the opportunity to observe the actual growth of a graph, so we can better understand how nodes and edges are added, e.g. whether the additions are linear or how many edges are formed among the existing nodes. This will provide an insight for the growth, and thus better guide future studies on complex networks.



(a) Node growth of P2P
(b) Node growth of HTTP
(c) Node growth of IM
(d) Node growth of Overall

Fig. 5. Node growth of flow graphs

We have analyzed the growth of the four flow graphs. The node growth processes over time are shown in Fig.5. The node growth over time is almost constant for the P2Pdownload applications and the overall traffic, which indicates that nodes do arrive at a constant rate when the associated traffic does not depend on human activities. The node growth of IM and HTTP vary with time, showing that there are busy hours when the applications are more active. The beginning parts of all the graphs have a curve shape showing nonlinearity. The reason of this inconsistency lies in the way the graph was constructed: the graph was built from an empty graph without any nodes or edges, and each flow created two new nodes and a new edge at the beginning. After a while, some flow records do not create new nodes and edges, since these flows are between nodes that are already in the graph. Only flow records with an IP address never shown before will create one or two nodes and an edge. So at the beginning, the growth of nodes and edges are faster than the later when the graph creation becomes stable. Indeed, the growth at an earlier time should contribute to the accumulative value of the later time. The actual flow graph should not be built from the blank. A flow record is not necessarily brand-new because there may be connections before the time point we start monitoring. So the actual growth should be like the latter part of the plot. In a real network, the actual rate of node increasing should vary with different applications



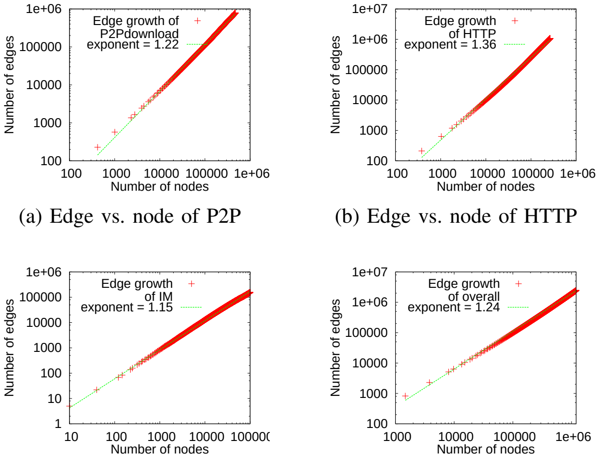
(a) Edge vs. node of P2P

(b) Edge vs. node of HTTP

(c) Edge vs. node of IM

(d) Edge vs. node of Overall

Fig. 6. The growth of edge vs. node



(a) Edge vs. node of P2P

(b) Edge vs. node of HTTP

(c) Edge vs. node of IM

(d) Edge vs. node of Overall

Fig. 7. The growth of edge vs. node

and time of the day, i.e. in busy hours, flows and thus nodes arrive faster. However, during the period of our records and thus the construction of graphs, the rates remain almost constant. We can see the node arrival rate of the overall graph is much larger than those of other graphs, which indicates many other types of applications exist besides the three we study.

While the rate of nodes added can vary with time, the edges added and their association with nodes can give us more insights of the growth of a graph. Though nodes are added to the graph almost constantly, it is not the case for the increase of edges. In Fig.6, we show the plots of the number of edges against the number of nodes as the graph grows, and they are clearly nonlinear. The up-bending curve means that edges are not added to the network in proportion to nodes added, but at a faster pace. We show the same plots in Fig.6 again in Fig.7, but with loglog scale. The straight lines clearly show that the edges are added super-linearly with the increase of nodes, i.e.:

$$e \sim n^{\alpha} \tag{7}$$

with α being 1.22 for P2Pdownload (Fig.7a), 1.36 for HTTP (Fig.7b), 1.15 for IM (Fig.7c) and 1.24 for the overall (Fig.7d). So for the flow graphs, with the number of nodes in the flow graph growing linearly with time, i.e. nodes are added at a constant rate, edges are added faster and faster. Besides the average degree, a complex network has many properties, for example, clustering coefficient, diameter, average path length, betweenness, etc. These properties are all impacted by the creation rate of new edges in the network. Therefore, the studies and observations on the edge growth process have a profound effect on the research of complex networks. The further studies on the other properties of complex networks will be left for our future work, and are beyond the scope of this paper.

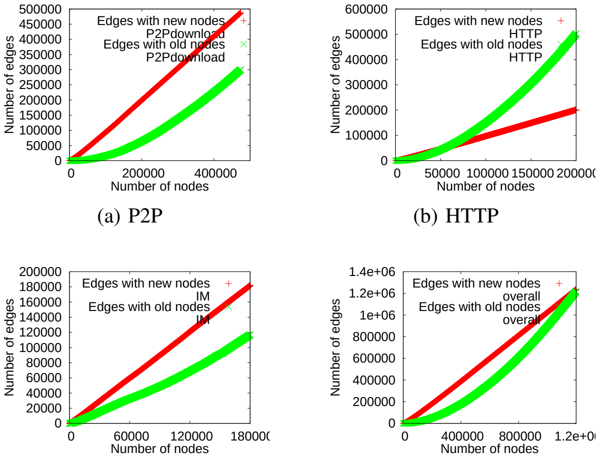
From Fig.6 and Fig.7, the nonlinearity is obvious for P2Pdownload, HTTP and overall traffic, but it is not clear for IM. The edge growth process of IM in Fig.6 roughly follows a straight line, and has the smallest exponent in Fig.7. The linear growth hypothesis of BA model is supported by some observations of human interactions, such as citations and social networks, which are similar to the IM application, but are different from machine-oriented applications such as P2Pdownload. This may imply that we should consider different models for different types of complex networks.

Most growth models of complex networks predict that the parameters of network grow slowly with the network size [2], i.e. the diameter and average path length of the network grow with the number of nodes n as $log(n)$ or $log(log(n))$. However there are observations that these properties actually decrease with the increase of n, i.e. network shrinks over time and growth [11]. The growth of our flow graphs reveals a possible reason for this: with a faster rate of edge creation over node increase within the network, the network becomes better connected, i.e. has a smaller diameter and average path length. According to this observation, the growing models of complex networks should not only consider the edges associated with new nodes, but also those between existing nodes.

Now it is clear that edges are added to the flow graph super-linearly to nodes added. Next, we try to answer the following questions: how many edges are added with the addition of new nodes and how many edges are created among the old ones. We classified each flow record and counted the number of edges associated with new nodes and old nodes separately. The results in Fig.8 show that the nonlinear part of the edge growth is clearly due to the edges formed between the old nodes, while the number of edges added in association with new nodes is linear with the growth of nodes. Moreover, the edges added between old nodes are much more than the edges coming with new nodes in the HTTP graph. This demonstrates that, for some applications, when the graph is large enough, the edges changing and forming between existing nodes will dominant the change of the network topology, and is thus much more important than the arrival process of new nodes.

C. *Preferential Attachment*

Preferential attachment (PA) is another key factor in BA model for the node degrees to achieve power law distributions



(a) P2P

(b) HTTP

(c) IM

(d) Overall

Fig. 8. Number of edges with new and old nodes

besides the incremental growth. PA means if $\pi(k,t)$ is the probability of an edge connecting to any node with k degree at time t, then for a larger k, $\pi(k,t)$ is larger. However, it is not clear that whether the relationship is linear. Some observations show linearity in one kind of graph [15], while some others show nonlinearity [16].

Our flow graphs are also useful for studying PA behavior. We follow the method used in [15], by defining a relative probability $R(k)$ as the relative probability of an edge connecting to a node with k degree, then the probability $\pi(k,t)$ of an edge connecting to any node with degree k can be represented as:

$$\pi(k,t) = R(k)p(k,t) \tag{8}$$

where $p(k,t)$ is the probability of a node with k degree at time t. We can use $n(k,t)/N(t)$ to estimate $p(k,t)$:

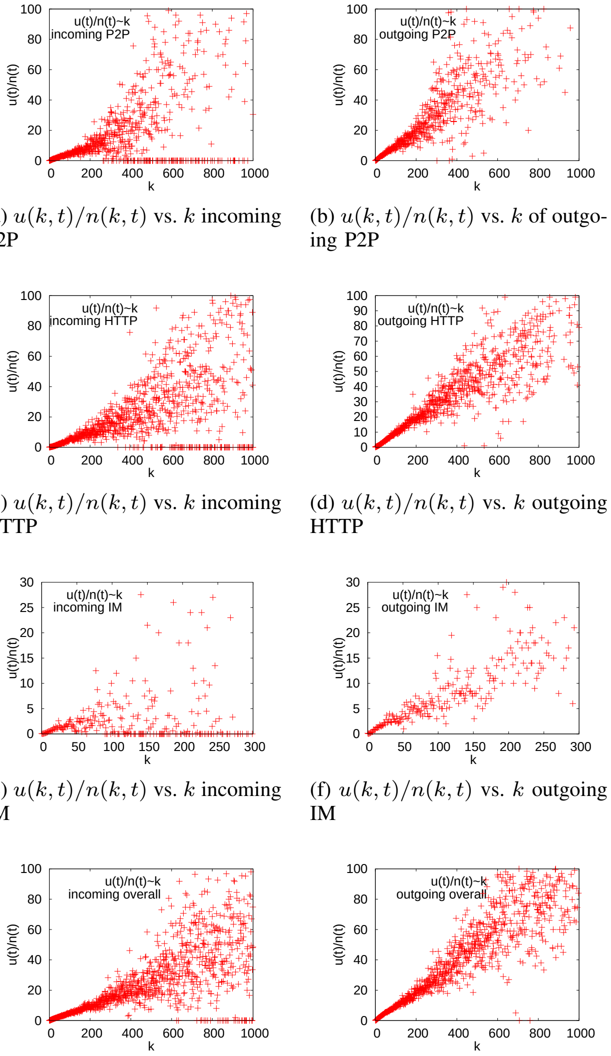
$$p(k,t) \approx n(k,t)/N(t) \tag{9}$$

where $n(k,t)$ is the number of nodes with degree k at time t, and $N(t)$ is the total number of nodes in the graph at t.

Choosing a time t at which the flow graph is sufficiently large, we retrieve $n(k,t)$ and then start counting newly arrived edges to acquire $u(k,t)$, which is the number of new edges connected to nodes of k degree. After a certain time interval dt, we stop counting and $u(k,t)$ can be used to estimate $\pi(k,t)$. By studying the relationship between $u(k,t)$ and $n(k,t)$, $R(k)$ can be estimated. If dt is small compared to t, this sample of $u(k,t)$ and $n(k,t)$ is accurate enough for the estimation.

We plot the histogram of $u(k,t)$ weighted by $1/n(k,t)$. If there is a linear preferential attachment, i.e. $R(k) \sim k$, we should see a straight line with its slope greater than 0 in the plot. If there is no preferential attachment, $R(k)$ should be a constant for all k values. In Fig.9, where the histogram of incoming and outgoing $u(k,t)/n(k,t)$ of the four flow graphs are shown, the relationship is clearly not constant. In order to make a better estimation of $R(k)$, we assume that $R(k)$ takes an exponential form, so we have:

$$R(k) \sim k^{\beta} \sim u(k,t)/n(k,t) \tag{10}$$



(a) $u(k,t)/n(k,t)$ vs. k incoming P2P

(b) $u(k,t)/n(k,t)$ vs. k of outgoing P2P

(c) $u(k,t)/n(k,t)$ vs. k incoming HTTP

(d) $u(k,t)/n(k,t)$ vs. k outgoing HTTP

(e) $u(k,t)/n(k,t)$ vs. k incoming IM

(f) $u(k,t)/n(k,t)$ vs. k outgoing IM

(g) $u(k,t)/n(k,t)$ vs. k incoming Overall

(h) $u(k,t)/n(k,t)$ vs. k of outgoing Overall

Fig. 9. $\pi(k,t)$ vs $n(k,t)$

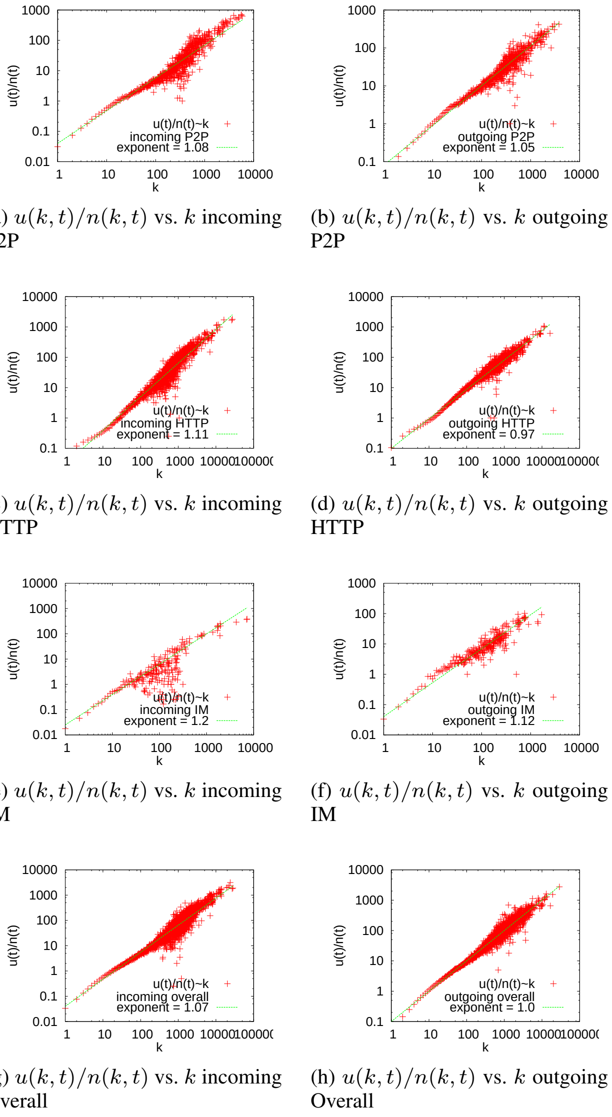
TABLE V
EXPONENTS OF PREFERENTIAL ATTACHMENT

	P2Pdownload	HTTP	IM	Overall
β_{in}	1.08	1.11	1.2	1.07
β_{out}	1.05	0.97	1.12	1.0

By plotting $u(k,t)/n(k,t)$ against k in a loglog plot (Fig.10), we can estimate the β. The values of β are summarized in Table V. All applications have β close to 1, showing preferential attachment is almost linear for both incoming and outgoing degrees.

V. CONCLUSIONS

In this work, we construct weighted flow graphs based on detailed Internet traffic flow records. The flow graphs reveal not only the structure or connections of different Internet sites, but also the way various sites are visited. By studying flow



(a) $u(k,t)/n(k,t)$ vs. k incoming P2P

(b) $u(k,t)/n(k,t)$ vs. k outgoing P2P

(c) $u(k,t)/n(k,t)$ vs. k incoming HTTP

(d) $u(k,t)/n(k,t)$ vs. k outgoing HTTP

(e) $u(k,t)/n(k,t)$ vs. k incoming IM

(f) $u(k,t)/n(k,t)$ vs. k outgoing IM

(g) $u(k,t)/n(k,t)$ vs. k incoming Overall

(h) $u(k,t)/n(k,t)$ vs. k outgoing Overall

Fig. 10. $\pi(k,t)$ vs. $n(k,t)$,loglog

graphs from the complex network point of view, we give out results on node degrees and their distributions of several types of applications, including P2Pdownload, HTTP, IM. Besides the power law behavior persisting in all applications, the average/maximum degrees and power law exponents show the differences in the nature of applications, which are helpful for both understanding the characteristics of Internet traffic and managing/engineering the Internet.

Taking advantage of the timing information of the flow records, we study the growth of flow graphs. We identify the nonlinearity of the growth of edges vs. nodes, and discover that many edges are formed between existing nodes. The number of edges added between existing nodes is larger than that associated with new nodes in some applications, which indicates that edges created between existing nodes are more important for studying the properties of a graph. Moreover, we found the existing connections are used repeatedly over time. This indicates that instead of only studying whether two nodes are connected, it may be more important to understand how they are connected, i.e. how many data are exchanged, how long the connection is active, etc. For this purpose, data exchanged between nodes are calculated as the strength of edges, and their distributions are also observed to exhibit power law behavior.

We also study the preferential attachment behavior based on the timing information of the records. Our observations confirm that PA exists, i.e., more edges are connected to a node with a higher degree, and the preference is almost linear. Our findings of linear preference and nonlinear growth of edges, especially the large number of edges formed between existing nodes are important in developing complex network models.

Our studies of different applications produce more comprehensive results. The parameter comparisons of IM and P2Pdownload clearly show the difference in traffic characteristics between human-oriented and machine-oriented applications, implying that the connection mechanism has a profound impact on the formation of complex networks, and more flexible models need to be investigated to capture the characteristics of different applications.

References

[1] M. E. J. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. USA, 98 (2001).

[2] R. Albert and A.-L. Barabsi, Statistical mechanics of complex networks, Reviews of Modern Physics 74, 47 (2002).

[3] A-L. Barabsi and R. Albert, Emergence of scaling in random networks, Science, vol. 286, pp. 509-512, Oct. 1999.

[4] D. J. Watts and S. H. Strogatz, Collective dynamics of small world networks, Nature, vol. 393, pp. 440-442, June 1998.

[5] J.-J. Pansiot and D Grad, On routes and multicast trees in the Internet. ACM SIGCOMM Computer Communication Review, 28(1): 41-50, January 1998.

[6] M. Faloutsos, P. Faloutsos, and C. Faloutsos, On power-law relationships of the internet topology, Computer Communications Rev., 29 (1999), pp. 251-262.

[7] P. Mahadevan, et al., The internet AS-level topology: three data sources and one definitive metric, Computer Communication Review 36, 17 (2006).

[8] R. Albert, H. Jeong and A.-L. Barabsi, Diameter of the World Wide Web, Nature, vol. 401, pp. 130-131, Sept. 1999.

[9] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Graph structure in the web, Computer Networks, 33 (2000), pp. 309-320.

[10] M. R. Meiss, F. Menczer and A. Vespignani, Structural analysis of behavioral networks from the Internet, 2008 J. Phys. A: Math. Theor. 41 224022.

[11] Jure Leskovec, Jon Kleinberg and Christos Faloutsos, Graphs over time: densification laws, shrinking diameters and possible explanations, Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD05), pp. 177-187, 2005.

[12] Amit A.Nanavati, et al., Analyzing the Structure and Evolution of Massive Telecom Graphs, IEEE Transactions on Knowledge and Data Engineering, Volume 20, May 2008.

[13] Krapivsky, P. L., S. Redner, and F. Leyvraz, Connectivity of Growing Random Networks, Phys. Rev. Lett. 85, 4629, 2000.

[14] Albert, R., and A.-L. Barabasi, Topology of Evolving Networks: Local Events and Universality, Phys. Rev. Lett. 85, 5234, 2000.

[15] M. E. J. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64, 025102(R) (2001).

[16] H. Jeong, Z. Neda, and A.-L. Barabsi, Measuring preferential attachment for evolving networks, Europhysics Letters 61, 567-572 (2003).

[17] M. E. J. Newman and Juyong Park, Why social networks are different from other types of networks, Phys. Rev. E 68, 036122 (2003)